Laaksonen, Salla-Maaria; Pääkkönen, Juho; Öhman, Emily

From hate speech recognition to happiness indexing: critical issues in datafication of emotion in text mining

Published in:
Handbook of Critical Studies of Artificial Intelligence

DOI:
10.4337/9781803928562.00064

Published: 01/11/2023

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Laaksonen, S.-M., Pääkkönen, J., & Öhman, E. (2023). From hate speech recognition to happiness indexing: critical issues in datafication of emotion in text mining. In S. Lindgren (Ed.), *Handbook of Critical Studies of Artificial Intelligence* Edward Elgar. https://doi.org/10.4337/9781803928562.00064

This is an author accepted manuscript version of a chapter forthcoming in the **Handbook of Critical Studies of Artificial Intelligence** by Edward Elgar Publishers. ISBN: 978 1 80392 855 5

Citation: Laaksonen, S-M., Pääkkönen, J. & Öhman, E. (2023). From hate speech recognition to happiness indexing: critical issues in datafication of emotion in text mining. In Lindgren, S. (ed.), Handbook of Critical Studies of Artificial Intelligence. Edward Elgar Publishers.

From hate speech recognition to happiness indexing: critical issues in datafication of emotion in text mining

Salla-Maaria Laaksonen[1] (0000-0003-3532-2387), Juho Pääkkönen[2] (0000-0002-0378-845X) and Emily Öhman[3] (0000-0003-1363-7361)

[1] University of Helsinki

[2] Aalto University

[3] Waseda University

Abstract

One prominent application of computational methods is the identification of affectivity and emotions in textual data, commonly known as sentiment analysis. In this chapter, we explore the datafication of affective language by focusing on operationalization and translation involved in the analysis processes behind common methods to identify affectivity or specific emotions in text. We draw examples from popular cases and from our own empirical studies that apply and develop sentiment and hate speech analysis. We suggest that sentiment analysis is a fruitful case for discussing the role of and the tensions involved in applying computational techniques in the automated analysis of meaning-laden phenomena. We highlight that any application of sentiment analysis techniques to investigate emotional expression in texts amounts to an effort of constructing sentiment measurements – a process essentially driven by judgments made by researchers in an attempt to reconcile diverging conventions and conceptions of good/proper research practices. (149)

Keywords: emotions, sentiment analysis, affectivity, hate speech, machine learning, datafication

Introduction

Sentiments and emotions are complex creatures that have been studied for millennia without much consensus as to their nature. A broad consensus lies, however, in their importance for social action. Scholars have argued for an 'affective turn' in social sciences and humanities to describe the growing interest in understanding the entanglement of social action and feelings (Clough & Halley, 2007; Wetherell, 2012). This scholarship has highlighted the importance of emotions and affect in organizing public discussions and social life. It has been suggested that the modern public sphere is best described as a space where publics are built as social formations through affective attunement (Papacharissi, 2015). Indeed, messages that gain traction in the public sphere are often affectively framed, and emotions are shown to motivate people and support the formation of new political groups and social action (e.g., Gould, 2010).

This rising interest in emotions has also had repercussions in the more technical and applied fields. In computer science, information systems research, and computational linguistics, emotions emerge as aspects of social life identifiable and measurable from natural language in computer-assisted or automated ways. The automated extraction and analysis of affective orientation in text is generally known as sentiment analysis (Pang & Lee, 2008). Sentiment analysis has emerged as a central procedure to understand the tone of public discussion. It is widely used in social media analytics and media research (Puschmann & Powell, 2018), while differing from traditional affect studies in its quantitative rather than qualitative approach. The methods used to analyze affectivity in texts draw largely on data mining, text representation and machine learning, commonly labeled as artificial intelligence tools in public discourse. While such tools are often presented as automated, intelligent and self-standing, their development and use involves human action and decisions at several stages of the process (e.g., Newlands, 2021; Rieder & Skop, 2021; Pääkkönen et al., 2020; Andrejevic, 2019). Furthermore, they require transforming the textual representations of emotions to quantified and datafied form, that is, into something that is recognizable and countable by machines (Beer, 2019).

In this chapter, we critically explore the datafication of affective language by focusing on the analysis processes behind a range of common sentiment analysis techniques. We discuss the underlying, often hidden choices related to datafication and operationalization of textual affect, and the various forms of human involvement when building technological systems to identify it. We argue that sentiment analysis provides an illuminating case for studying tensions in the use of artificial intelligence applications and computational technologies to understand complex social phenomena.

Tools to identify textual sentiments have become increasingly common in real-life contexts from social media monitoring to customer insights. As such, sentiment analysis has gained a somewhat institutionalized position as an indicator of the firm's current stance (Etter et al., 2018), or as a way to infer and predict social phenomena from public mood to voter preferences (Bollen et al., 2009; Oliveira et al. 2017). Further, it is used by the social media platforms themselves to understand the discussion flowing on their services and to aid in content moderation, particularly when removing unwanted content such as hate speech. Yet, to a regular user, the results of sentiment analysis are typically conveyed through a red-and-green diagram on an analytics platform – a representational format that hides potentially complex analysis processes and imbues the results with ostensible simplicity and mathematical truthfulness (Kennedy et al., 2016). Social media platforms, then again, present such endeavors as triumphs of technology: "Facebook Pulls 22.5 Million Hate Speech Posts in Quarter", or "YouTube removes more than 100,000 videos for violating its hate speech policy". What actually happens behind these diagrams, large numbers and success-reporting headlines, however, is rarely disclosed. As users of commercial platforms, we only live with the deliverables and decisions produced by these systems (Brown, 2018).

In what follows, we use our own computational projects and field work (e.g., Öhman, 2021a; Laaksonen et al., 2020; Haapoja et al. 2021) as examples to discuss and critically reflect on the processes of datafying textual expressions of affect from contested definitions to quantified algorithmic probabilities. We first discuss the difficulties in defining and operationalizing emotions as a multidisciplinary concept, and then discuss the datafication of emotions to commensurable digits. Next, we explain the validation processes typically involved in sentiment analysis and then move on to describe the hidden role of humans throughout the process. Through these explorations, the datafication of emotions for computational purposes emerges as a series of transformations where the phenomenon known to be contextual, embodied and complex, becomes reduced to simple numbers and compelling diagrams. Therefore, it becomes an affective

object that is commensurable and thus seemingly controllable for the society that increasingly strives for rationality and technological control.

Operationalization: what is affect/emotion/feeling/sentiment?

Sentiment analysis is often used as an umbrella term for many different types of investigations of affect and emotion in text (Munezero et al., 2014). In certain fields *sentiment analysis* is strictly the classification of polarity (positive—negative), sometimes including neutral and/or ambiguous. In other fields, any emotion detection and multi-label emotion classification is referred to as sentiment analysis. The main aim across these approaches is to identify emotions, which requires clear definitions. However, as emotions are studied from many different perspectives ranging from philosophical or psychological, to physiological, disciplines have developed slightly differing definitions of emotions as well as distinctions between opinions, sentiments, feelings, emotions, and affect (Izard, 2007; Nabi, 2010). In broader terms, we can say that *sentiments* are longer lasting and more general than *emotions*, *affect* is prepersonal, *feelings* personal, and *emotions* social (Shouse, 2005). Therefore *affect* is the most subconscious, followed by *feeling*, with *emotion* being involuntary but conscious, and *sentiment* and *opinion* the most conscious and prompted by feelings and emotions (Öhman, 2021a).

Many emotion models aim to organize emotions in various ways. In principle, two basic models of emotion can be distinguished: dimensional and discrete (Nabi, 2010). Dimensional models see emotions as a motivational state characterized by two or three broad affective dimensions, whereas discrete emotion perspectives identify emotional states by the unique set of cognitive appraisals, or thought patterns, underlying them. For example, the dimensional Russell's circumplex model, typically used in empirical psychology and psychophysiology, maps the emotional space with two orthogonal dimensions: emotional valence and emotional arousal (Russell, 1980). Discrete models, then again, propose that there are a small set of discrete emotions. Tomkins (1962), for example, suggested that there are nine basic affects: interest, enjoyment, surprise, distress, fear, anger, shame, dissmell and disgust, each based on biological affective responses. Yet, there is a surprising lack of consensus even among the basic classifications of emotions.

Behind the definitions lie more fundamental epistemological differences between positivist and interpretivist approaches, and the general question of what is quantifiable. Positivist approaches aim to measure emotions or causal factors that influence them. One critique of these approaches is that some feelings, such as hope, are more 'intellectual' (Averill, 1996), and cannot be easily simplified to quantifiable items. By contrast, interpretivist approaches focus on descriptive and processual aspects of emotions, and their discursive representation and production. This approach is more traditional in social sciences, exploring for instance how cultural context affects feelings and how we describe them, and the norms and values that place expectations on our emotional behavior. A central argument is that as language is the principal tool for meaning-making and communication, it is also the only means by which we can think or feel anything (Gerth & Mills, 1954). That is, emotions are discursively defined by the way they are described in language. That is why emotions have no life of their own independent of language and culture – a notion on which the interpretivist approach criticizes the positivist accounts.

The fact that the debate on the exact terminology and nature of emotions is ongoing is not necessarily a drawback. The different views of affect and emotion complement each other and create a wider base for understanding affective language (Alm, 2010). However, the multiplicity of

conceptions entails that the task of quantifying and measuring emotion expressions becomes a methodologically complex task that cannot be reduced to simple technical criteria. This makes sentiment analysis and its cognate approaches particularly fruitful cases for discussing the role and tensions involved in applying computational techniques in the analysis of meaning-laden phenomena. For most practical purposes, it makes sense to look at specific emotions rather than binary polarity. Although we can map many emotions onto positive and negative, knowing emotion over sentiment produces a more fine-grained overview of the data. There are two main approaches to sentiment analysis: data-driven and lexicon-based, with the addition of hybrid approaches that use lexicons in conjunction with data-driven methods.

Data-driven approaches refer mainly to machine learning. Classification algorithms are often used together with state-of-the-art language models such as BERT (Devlin et al., 2018). The one thing these approaches have in common is the need for annotated training data; the more data, the more precise the model will be and the more expensive the training will be too. Lexicon-based approaches, on the other hand, use predefined dictionaries that define the words associated or evoked by each sentiment (Taboada et al., 2011). It is possible to create these dictionaries automatically by associating, for example, a specific emoticon, emoji, or hashtag with a correlating emotion (De Choudhury et al., 2012), but typically human annotators are used. Although this criticism is more or less true depending on the specific approach, lexicons offer transparency and re-usability that machine learning approaches have difficulty achieving. Lexicons are less domain-dependent (Taboada et al., 2011) and require less CPU/GPU cost. Intensification, valence-shifters, and negations can all be incorporated into a lexicon-based approach increasing context-awareness. Further, in terms of this chapter, their critical analysis also provides insights into the processes of generating training data and numerical representations which are both essential for machine learning approaches as well.

Recognition and calculations

Sentiment analysis aims to identify emotions in text, and as explained above, utilizes simple or more sophisticated text mining methods to conduct this task. Most automated methods start by building mathematical representations of words in texts, such as frequency matrices or word vectors in n-dimensional space. In language technology, it has been suggested that existing systems will progress from syntactics (bag-of-words) to semantics (bag-of-concepts) to pragmatics (bag-of-narratives) in the coming decades (Cambria et al., 2017). Currently, however, most forms of applied sentiment detection rely, in one way or another, on word lists, bag-of-word approaches, or ngrams (e.g., Acheampong et al., 2020, Chen et al., 2012; Theocharis et al., 2020). This means counting the frequency of specific words or word combinations extracted from a document. These approaches are criticized for simplicity, ignoring context and even negation and valence shifters. Some more sophisticated methods utilize bag-of-word vectors combined with word dependencies to identify syntactic grammatical relationships in sentences (Burnap & Williams, 2015), semantic word embeddings (Badjatiya et al., 2017), or neural networks (Al-Makhadmeh & Tolba, 2019; Relia et al., 2019). Although such models have been called “stochastic parrots” (Bender et al., 2021), they emulate human understanding reasonably well—but in the end, even the most advanced deep learning algorithm counts words and produces previously seen phrases based on statistical likelihoods of co-occurrence.

Affectivity, however, is much more than words. Research on emotions has long recognized that textual expressions are not necessarily directly connected to authors' felt emotions (e.g., Wetherell, 2012). Similarly, while words are indicative of a specific emotion, emotions cannot be reduced to specific textual formations or combinations of words. The actual sentiment or affective

tone of a particular message relies immensely on the final form of the expression. Context-aware systems, such as word embeddings, should enable fine-grained understanding of word contexts and semantics. Yet, neither have access to actual feelings or intentions of the author, or even feelings expressed by the text—the words and contexts in the texts merely statistically tend to express a certain feeling or emotion. Despite this, some researchers (e.g., Pennebaker et al., 2015) advocate the possibility to deduce a writer's emotional state and even psychometrics based on their word choices.

Despite these discrepancies, automated text mining methods require the transformation of textual data into word representations that can be expressed in a numerical format and then processed using mathematical operations. This quantification and simplification was highlighted in the hate speech detection project described by Laaksonen and colleagues (2020), where the main goal of the project essentially turned out to be quantification of affect using a rudimentary scale from 0 to 3 to annotate the severity of hate speech. The authors describe working with the spreadsheets of data as a blunt moment of quantification: turning message content and meaning to a single digit, a figure of anticipation (Mackenzie, 2013), stripping off nuances in verbal expression. Likewise, lexicon-based sentiment analysis models rely on sentiment weights given to words to indicate the strength of their association with a specific emotion. A recent project by the authors of this chapter utilized the lexicon developed by Öhman (2022), fine-tuned to match a novel dataset by examining the 3,000 most frequent words in the dataset to add missing emotions, edit associated emotions, and remove incorrect associations. Editing associations meant tweaking the number that defines the weight of the word-emotion association. In this process, the researchers ended up numerically defining the weight between, for example, ‘pandemic’ and ‘fear’, without knowing the context in which the word will appear.

Quantification inevitably flattens the data and results in loss of variety in expressions, thus performing inevitable oversimplification and formalization of a cultural phenomenon such as emotions. It could be argued that this is precisely what makes algorithms and algorithmic systems powerful; their ability to perform abstraction (Pasquinelli, 2015, cited in Mackenzie, 2017). We argue, however, that the translations involved in datafying affective language go beyond abstraction; due to the nature of the methods used to extract information from text, sentiment values produced by sentiment analysis are objects produced by and for the automated methods used to measure them. As Espeland and Stevens (2008, p. 411) note, quantification is "work that makes other work possible", by establishing *shared* practices for representing and measuring interests in terms of numeric values that yield for further analytical processes. This construction of sentiment becomes evident when detecting complex forms of affective language like hate speech: are we measuring hateful words, presence of negative emotions, framings, or accusations and calls for action targeted to certain groups of people, as postulated in international agreements on hate speech? Studies of data science have shown that efforts to establish trust in the results of computational knowledge production cannot be understood merely by reference to technical criteria, but instead involve social processes of negotiation through which diverging interpretations are reconciled (Passi & Jackson, 2018). In such negotiations, the challenge is to coordinate approaches to data work and modeling with broader socially established conventions of acceptable analysis – which in loosely interlinked communities such as 'artificial intelligence research' can vary widely with respect to disciplinary background and training (Forsythe, 1993). These "human" issues in computational analysis become especially visible in the processes of modeling result validation and training data generation, which we discuss next.

Validation and novel contexts

As mentioned earlier, the theories of emotion that sentiment analysis is based on are psychological theories meant to describe human behavior “in the wild”. However, when the target of the investigation is text, it could be argued that the output is more deliberate compared to social interactions or psychological phenomena. For example, in psychology, Plutchik (1980) considers anger as a positive emotion because it is an active one, but it makes little sense to classify anger as positive when using applied sentiment analysis to detect hate speech or monitor customer feedback. Current psychological categorizations of emotions thus often are inadequate in describing online communication. Cowen and Keltner (2017) and Demszky et al. (2020) have attempted to rectify this by developing novel emotion categorization systems. However, here we stumble upon the limitations of machine learning algorithms and the overlapping nature of emotions. Machine learning classification works best when categories are as disjoint as possible. Emotions are, however, not disjoint, and the same physiological response can be expressed differently across cultures and linguistic environments.

State-of-the-art solutions are virtually all data-driven, although this statement is debatable due to different validation techniques used with data-driven and lexicon-based approaches. When using a machine-learning based approach to sentiment analysis, validation is built-in to the process. The entire premise is to have your data split into training and testing subsets to be able to calculate f1-scores (or similar) to determine the accuracy of the model (Hastie et al., 2009). The unit which is tested is the same as the unit that is being tested against, making the validation process straightforward. The advantage of machine learning is that models can distinguish between contexts, allowing polysemous words to be categorized as different emotions based on context. Nevertheless, it is important to recognize that a model trained using a certain dataset is valid only in relation to that particular data. This observation highlights that validation in data-driven analysis is based on upstream processes such as data selection, curation, and annotation, which involve laborious manual interpretive work.

When using lexicon-based solutions, we often face a mismatch between the units that are being compared. In lexicon-based approaches only the test unit is manually annotated by humans, and therefore, the compared units nor their annotations are identical. In essence, with most lexicon-based approaches we are calculating a composite score of emotions detected and normalized for word count. A single unit of data is also often much larger than with machine learning approaches (speech vs. product review). Hence, the human annotation is difficult to match to the output of the lexicon-based model as the unit of annotation is different, and a phrase where the computer finds several word matches a human is more likely to assign a single label (Öhman, 2021b). If accuracy measurements in machine learning are comparing apples to apples, in lexicon-based approaches validation requires the comparison of oranges to bread baskets.

Despite these difficulties in validating lexicon-based approaches, the evidential weight of analysis tends in practice (e.g. in review processes) to be ascribed heavily on validation, which can give rise to procedures that are neither helpful nor informative, even with respect to the actual accuracy of the model. On the other hand, especially in the light of the popularity of blackbox tools in some interdisciplinary fields, e.g. LIWC or Vader-sentiment, the results are often accepted without much validation, even though the process is no more robust than other lexicon-based approaches. The allure of off-the shelf products in their simplicity and ability to provide simple representations of emotions in text, should not due to in-field tradition be exempt from robust validation methods. Incorporating a qualitative component to sentiment analysis is one way to enhance contextual validity.

Training data generation as contested negotiation

Discussions of computational text analysis in the social sciences and humanities have long emphasized the crucial role of discretionary human judgment and interpretive engagement in modeling processes. While computational approaches are often associated with notions such as data-drivenness, general recognition among scholars is that all modeling procedures depend on a series of preliminary decisions and negotiated analysis practices that cannot easily be automated (Grimmer & Stewart, 2013). Articulating the role of human discretionary work in automated analysis approaches becomes increasingly difficult once they become established methodological protocols that researchers can cite as authoritative academic references (Betti & van den Berg, 2016). Indeed, the magic of machine learning is that it is easy to follow the programmatic knowledge production practices in the field, not only to organize data, but also the relationships between humans and machines in the process (Mackenzie, 2017).

This interwoven nature and the following emergence of *constructed* sentiment measurements is well demonstrated by data-driven approaches to sentiment analysis, which draw on variants of supervised machine learning. As noted above, data-driven sentiment analysis depends on training data, fed to the learning algorithm as a datafied definition of the desired target. It is well known that the quality and content of training data highly affects the model (e.g., Hastie et al., 2009; Mackenzie, 2017). When choosing the data set, we give cues to the machine learning algorithm as to what kind of content we are looking for. These cues depend on first, the availability of data and second, on our ability to select reliable, representative data. Biases caused by training data are sometimes rather obvious in existing systems. For example, Google Jigsaw Perspective API, a state-of-the-art model for toxic language detection (see Rieder & Skop, 2021), has been accused of giving higher toxicity scores to sentences that mention women than men (Jigsaw, 2018). Such differences are due to the over-representation of certain classes in the training data that the system is built on. Unless carefully balanced, any real-life dataset contains more toxic comments concerning regularly targeted groups, so the evaluation of toxicity becomes attached to words that should represent the “neutral context.” It is important to note that such biases are difficult to anticipate before being exposed through audits or scandals; however more transparent documentation practices for datasets have been suggested as a potential solution (e.g., Gebru et al., 2021).

The difficulties of identifying and rectifying biases recur throughout the process of training data generation. In data-driven sentiment analysis, the standard approach is to generate training data by manually annotating a smaller set of messages (e.g., a random sample), selected from the full dataset to be classified. Typically, multiple human annotators classify the data individually before a final label is assigned, based on comparison and discussion of the individual annotations. In order to reduce noise, the annotators first have to reach agreement upon the labeling criteria that guide annotations. Noisy annotations are those labels which annotators have difficulty agreeing upon, often revealed by inter-annotator agreement/reliability calculations (Krippendorff, 2003). However, as there is rarely a single "correct" emotion label for a specific text, it is understandable that humans seldom agree on annotations. Therefore choices need to be made on how to deal with discrepancies: whether to remove all contested annotations, relabel them as neutral, use majority voting, or defer final judgment to an expert annotator? Moreover, certain emotions are more difficult to annotate in text. For example, “surprise” is often highly contextual (Alm & Sproat, 2005). Likewise, annotating hate speech has been shown to be a difficult task for humans (Davidson et al., 2017; Laaksonen et al., 2020). The need for iterative negotiation and revision of annotation criteria makes training data generation a costly procedure, which often does not yield simple one-off solutions. The quality of annotations is directly related to the accuracy of modeling

outcomes. However, whether reducing noise to achieve higher accuracies results in more accurate emotion detection in terms of what emotions are actually expressed in texts, is highly debatable.

The need for interpretative engagement at the outset of sentiment analysis throws into relief the constructed nature of the resulting measurements. While the act of assigning numerical values to objects can amount to a simple practice of arbitrary "marking" (Espeland & Stevens 2008), the establishment of common analytical practices around numeric representations is a social achievement in coordination. For instance, as Laaksonen et al. (2020) showed in their analysis of a hate speech identification project, even after several rounds of coding and revisiting the initial definitions, the resulting definition of hate speech retained unclarities and inconsistencies when applied in practice. These observations align with academic reports on hate speech recognition, which highlight difficulties in separating hate speech from other types of offensive language (e.g., Davidson et al., 2017). Indeed, Pàmies et al. (2020) showed that offensive social media messages included more emotion-associated words than non-offensive messages—both positive and negative.

More generally, annotation decisions are dependent on annotators' previous knowledge, in regard to the classifying task, context, and cultural connotations (Waseem & Hovy, 2016; Davidson et al., 2017). Observations of annotation tasks have shown that even with binary or ternary classification schemes, human annotators agree only 70-80% of the time and the more categories, the harder it becomes to reach agreement (Bayerl et al., 2011; Öhman, 2021a). The question then becomes: how can a machine learning algorithm learn from human annotated data where most of the annotations are contested? Model performance is highly dependent on the quality of the annotations, but even with the best trained annotators and most carefully selected and defined categories, if humans – who are the supposed experts at natural language understanding – cannot tell emotion categories apart, how can we expect computers to succeed?

Conclusions and implications

Text mining emotions is a challenging technical endeavor—but perhaps consequently, it represents a type of societal issue many actors hope to solve with technology. As discussed in this chapter, these solutions hinge on complex processes of datafication and quantification of emotions and affective language. Our discussion highlights that utilizing automated, AI-assisted methods typically include several undisclosed steps of human judgment, from training data annotation to validation to questions of model selection and research problem formulation. From this perspective, emotion detection emerges as a process that combines interpretative validation and mathematical operations. The research design constellation and choices made along the way define whether the extracted sentiments are emotions, affective states, or something else. The development of the methods and excitement related to novel technological opportunities steers the process: state of the art methods are typically wanted in academic publications and eagerly tested in applied contexts despite the escaping explainability of their operation. Indeed, off-the-shelf models from LIWC to BERT are conveniently packaged and easy to use without extensive technical knowledge.

Throughout this chapter we have characterized automated emotion detection as a *process of constructing measurements of sentiment expressions* in text. This view implies that the results of sentiment analysis should not be evaluated solely on the basis of certain pregiven metrics that evaluate, for example, model validity in terms of accuracy on a dataset of pre-annotated messages. Rather, model evaluation in general and validity assessment in particular should be considered as a process that cuts through the whole process of research design and

understanding of the research problem at hand. Ultimately, questions about the validity of measurements of complex meaning-expressions should depend on criteria such as collective assessment through reflexive debate about definitions. Such criteria can be difficult to uphold and reconcile with the demands and diverging conventions imposed on reports of computational research by diverging publication venues and communities of practice. We contend that sentiment analysis is a fruitful case through which to reflect on these processes, given that these families of methods stand at the juxtaposition of stringent technical evaluation practices on the one hand, and the aim of investigating meaning-laden concepts with highly varying interpretations across contexts on the other.

Technologies for detecting emotions are increasingly used in the industry, for example to monitor and moderate online discussions, and hence will reshape our communication environments and our society in the future (Brown, 2018). Applications of hate speech detection used by social media platforms or NGO monitoring projects are a prime example of the applied versions of emotion detection. Other, more playful yet convincingly presented applications are attempts to measure public mood and cross-national happiness using sentiment analysis of social media data (Bollen, 2009; Kamer, 2010; Dodds et al., 2011). Particularly in public discourse such projects are commonly presented and marketed through the magical aura of artificial intelligence (Newlands, 2021). Yet we know that algorithmic systems rarely perform their tasks perfectly when dealing with complex language data (Grimmer & Stewart, 2013), and that development and use of AI, being a socio-technical system, can never be separated from the social context (Elish & boyd, 2018). We argue that a better understanding of the capabilities and limitations of these technologies is needed, and action research is one way to generate that knowledge and pursue interventions (Kennedy et al., 2015; Laaksonen et al., 2020; Rieder & Skop, 2021).

References:

Acheampong, F., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports, 2*(7), e12189.

Alm, C. (2010, July). Characteristics of high agreement affect annotation in text. In *Proceedings of the fourth linguistic annotation workshop* (pp. 118-122).

Alm, C., & Sproat, R. (2005, October). Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 668-674). Springer, Berlin.

Andrejevic, M. (2019). Automating Surveillance. Surveillance & Society, 17(1/2), 7–13.

Averill, J. (1996). Intellectual emotions. In Harre, R. & Parrot, W. G. (Eds.) *The emotions: Social, cultural and biological dimensions* (pp. 24-38). Sage.

Bayerl, P., & Paul, K. I. (2011). What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics, 37*(4), 699-725.

Bender, E.., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

Beer, D. (2019). *The Data Gaze: Capitalism, Power and Perception*. SAGE.

Betti, A. & van den Berg, H. (2016). Towards a Computational History of Ideas. In *CEUR workshop proceedings* 1681.

Bollen, J., Pepe, A., & Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *arXiv preprint.* http://arxiv.org/abs/0911.1583

Brown, A. (2018). What is so special about online (as compared to offline) hate speech? *Ethnicities 18*, 297–326.

Burnap, P., & Williams, M. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet, 7*(2), 223–242.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In A practical guide to sentiment analysis (pp. 1-10). Springer.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In *Proceedings of the Fourth ASE/IEEE International Conference on Social Computing* (SocialCom 2012), September 3–6, Amsterdam.

Clough, P. & Halley, J. (2007). *The Affective Turn: Theorizing the Social.* Duke University Press.

Cowen, A., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences, 114(38), E7900-E7909.

De Choudhury, M., Gamon, M., & Counts, S. (2012). Happy, nervous or surprised? Classification of human affective states in social media. *ICWSM 2012*, 435–438.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. arXiv preprint. arXiv:2005.00547.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dodds, P., Harris, K., Kloumann, I., Bliss, C., & Danforth, C. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE, 6*(12), e26752. https://doi.org/10.1371/journal.pone.0026752

Elish, M., & boyd, D. (2018). Situating methods in the magic of Big Data and AI. *Communication Monographs*, *85*(1), 57–80.

Etter, M., Colleoni, E., Illia, L., Meggiorin, K., & D'Eugenio, A. (2018). Measuring Organizational Legitimacy in Social Media: Assessing Citizens' Judgments With Sentiment Analysis. *Business and Society*, *57*(1), 60–97.

Espeland, W. & Stevens, M. (2008). A Sociology of Quantification. *European Journal of Sociology, 49*(3), 401–436.

Forsythe, D. (1993). Engineering Knowledge: The Construction of Knowledge in Artificial Intelligence. *Social Studies of Science* 23(3), 445–447.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

Gerth, H., & Mills, C. W. (1954). *Character and social structure*. Taylor & Francis.

Gould, D. (2010). On affect and protest. In: Staiger, J., Cvetkovich, A., Reynolds, A. (Eds.), *Political emotions* (pp. 18-44). Routledge

Grimmer, J. & Stewart, B. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3), 267–297.

Haapoja, J., Laaksonen, S-M., & Lampinen, A. (2020). Gaming Algorithmic Hate-Speech Detection: Stakes, Parties, and Moves. *Social Media and Society, 6*(2).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Izard, C. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science, 2*, 260-280.

Jigsaw (2018). Unintended Bias and Names of Frequently Targeted Groups. Blog post on the False Positive/Medium. Available: https://medium.com/the-false-positive/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23 (accessed August 15, 2022).

Kennedy, H., Hill, R., Aiello, G. & Allen, W. (2016). The work that visualisation conventions do. *Information, Communication & Society* 19(6), 715–735.

Kennedy, H., Moss, G., Birchall, C., & Moshonas, S. (2015). Balancing the potential and problems of digital methods through action research. *Information Communication and Society, 18*(2), 172–186.

Kramer, A. (2010). An unobtrusive behavioral model of “gross national happiness.” *Conference on Human Factors in Computing Systems Proceedings*, 1, 287–290.

Krippendorff, K. (2003). *Content Analysis: An Introduction to Its Methodology*. Sage.

Laaksonen, S-M., Haapoja, J., Kinnunen, T., Nelimarkka, M., & Pöyhtäri, R. (2020). The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Frontiers in Big Data, 3*(February), 1–16.

Mackenzie, A. (2013). Programming subjects in the regime of anticipation: software studies and subjectivity. *Subjectivity 6*, 391–405.

Mackenzie, A. (2017). *Machine Learners: Archaeology of Data Practice.* MIT Press.

Munezero, M., Montero, C., Sutinen, E., & Pajunen, J. (2014). Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing, 5*(2), 101-111.

Nabi, R.. (2010). The Case for Emphasizing Discrete Emotions in Communication Research. *Communication Monographs, 77*(2), 153–159.

Newlands, G. (2021). Lifting the curtain: Strategic visibility of human labour in AI-as-a-Service. *Big Data & Society, 8*(1).

Oliveira, D., Bermejo, P. de S., & dos Santos, P. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology and Politics*, *14*(1), 34–45.

Pàmies, M., Öhman, E., Kajava, K., & Tiedemann, J. (2020, December). LT@ Helsinki at SemEval-2020 Task 12: Multilingual or Language-specific BERT?. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1569-1575).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.

Papacharissi, Z. (2015). *Affective publics: Sentiment, technology, and politics*. Oxford University Press.

Passi, S. & Jackson, S. (2018). Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction, 2* (CSCW).

Pennebaker, J., Boyd, R., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015.* University of Texas at Austin.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion* (pp. 3-33). Academic press.

Puschmann, C., & Powell, A. (2018). Turning Words Into Consumer Preferences: How Sentiment Analysis Is Framed in Research and the News Media. *Social Media + Society*, *4*(3).

Pääkkönen, J., Laaksonen, S-M. & Jauho, M. (2020). Credibility by automation: Expectations of future knowledge production in social media analytics. *Convergence* 26(4), 790–807.

Rieder, B., & Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of Perspective API. *Big Data and Society, 8*(2).

Russell, J., (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6), 1161-1178.

Shivhare, S., & Khethawat, S. (2012). Emotion detection from text. *arXiv preprint* arXiv:1205.4944.

Shouse, E. (2005). Feeling, emotion, affect. *M/c journal, 8*(6).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics, 37*(2), 267-307.

Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. (2020). The Dynamics of Political Incivility on Twitter. *SAGE Open, 10*(2).

Tomkins, S. (1962). *Affect imagery consciousness*. Springer.

Wetherell, M. (2012). *Affect and emotion: A new social science understanding*. Sage.

Öhman, E. (2021a). The Language of Emotions: Building and Applying Computational Methods for Emotion Detection for English and Beyond. Doctoral Dissertation, University of Helsinki.

Öhman, E. (2021b). The Validity of Lexicon-based Emotion Analysis in Interdisciplinary Research. In *Proceedings of NLP4DH @ ICON'21*. ACL Anthology.

Öhman, E. (2022). SELF & FEIL: Emotion Lexicons for Finnish. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*. CEUR.